Motor Trend Data Set: Automatic vs Manual Transmission for Fuel Consumption

Liubov Gryaznova

January 28, 2016

Executive Summary

In this project we examine the `mtcars` data set which originates from Motor Trend (US magazine) car road tests comprising fuel consumption and 10 aspects of automobile design and performance for 32 automobiles (1973–74 models). We explore the relationship between a set of variables and miles per gallon (MPG) as an outcome paying particular attention to the type of transmission.

We performed exploratory data analysis and fitted multiple linear models to select the most suitable one. Though at a first glance manual transmission results in more miles per gallon whereas automatic transmission tends to be more fuel consuming, we cannot prove it from the data given in the dataset. Variables are highly correlated with each other and some of them (e.g. weight of a car or horse power) have a greater influence on MPG than transmission.

Exploratory Data Analysis

We load the data set and look at its structure to work out further strategy. We will use `mpg` variable (miles per US gallon) as the outcome and `am` variable (transmission) as our first predictor since we are interested in the effect of the transmission type. Since `am` has only two discrete values we will treat it as a factor variable later in our modeling. First we make a boxplot to visualize if there is a relationship between transmission and fuel consumption (see figure 1 in the Appendix). According to the plot manual transmission demonstrates more MPG. So we run a t-test to check this statement and get very low p-value (<0.05) which confirms the difference between two groups. If we run the test with `alternative = "greater"` option we will also confirm our alternative hypothesis.

```
# t-test
auto <- mtcars$mpg[mtcars$am == 0]; manual <- mtcars$mpg[mtcars$am == 1]
c(t.test(manual, auto, alternative = "two.sided")$p.value,
  t.test(manual, auto, alternative = "greater")$p.value)
```

```
## [1] 0.0013736383 0.0006868192
```

Simple Linear Regression Model

We start with a simple model using only `mpg` as the outcome and `am` as a predictor. We get $R^2 = 0.3598$ which means that the model explains only 35.98% of the variation. Coefficients table demonstrates that the difference in means is statistically significant, so on average the MPG value for manual transmission is 7.245 greater than the respected value for the automatic transmission.

```
fit <- lm(mpg ~ factor(am), data = mtcars)
summary(fit)$r.squared; summary(fit)$coef
```

```
## [1] 0.3597989
```

```
##              Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 17.147368   1.124603 15.247492 1.133983e-15
## factor(am)1  7.244939   1.764422  4.106127 2.850207e-04
```

Multivariable Linear Regression Model

Our simple model does not explain much of variation as we have other predictor variables that might affect MPG. To get a better result we will try to fit an appropriate multivariable linear regression model. First we look at a model containing all the variables and find out that in this case none of predictors are statistically significant (according to the p-value). Therefore we reject this model and look at the correlation matrix plot (see figure 2 in the Appendix) to find appropriate variables to include in our model. We see that many variables are heavily correlated with each other which makes it difficult to fit a model. So we choose variables which seem to have more effect on MPG but not correlated heavily with each other and `am` not to inflate the variance. We add those variables one by one and check if they suit the model. We look at R^2 to see the amount of variation explained, at p-values to maintain significance and at residual plots which should not have any patterns.

Adding variables demonstrates that `am` tends to lose its significance in presence of confounders. We fit `fitp1` model by including horse power and `fitp2` model by including weight and quarter mile time in addition to transmission. Both models maintain `am` significance and explain much of variation (78.2% for `fitp1` and 84.97% for `fitp2`). Thus the second model shows better R^2 value, however the first one provides less VIF and a better residuals vs fitted values plot (see figures 3 and 4 in the Appendix). Since we cannot assure that there is no clear pattern in the `fitp2` residual plot we prefer `fitp1` model at the moment. Nevertheless we would like to improve R^2 value as well, so we try an automated approach.

```
fitp1 <- lm(mpg ~ factor(am) + hp, data = mtcars)
summary(fitp1)$coef
```

```
##               Estimate  Std. Error   t value     Pr(>|t|)
## (Intercept) 26.5849137 1.425094292 18.654845 1.073954e-17
## factor(am)1  5.2770853 1.079540576  4.888270 3.460318e-05
## hp          -0.0588878 0.007856745 -7.495191 2.920375e-08
```

We use `stepAIC` function from `MASS` module to find out the best model in an automated way. To make the function work correctly we transform `am` (transmission), `cyl` (number of cylinders), `vs` (type of engine, V or straight), `gear` (number of forward gears), and `carb` (number of carburetors) into factor variables. The model `fitstep` proposed by `stepAIC` includes transmission and number of cylinders (both as factors), as well as weight and gross horsepower. The choice of variables also seems quite reasonable due to the correlation table and common sense. It is a highly predictive model explaining 86.59% of variation. Here manual transmission demonstrates greater MPG value (the increase in MPG compared to automatic transmission is 1.809); however, p-value of this coefficient is high so it cannot be called statistically significant. VIF and residual plots (see figure 5 in the Appendix) are suitable so the model can be considered appropriate.

```
fitstep <- lm(mpg ~ factor(am) + factor(cyl) + hp + wt, data = mtcars)
summary(fitstep)$coef
```

```
##                  Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)  33.70832390 2.60488618 12.940421 7.733392e-13
## factor(am)1   1.80921138 1.39630450  1.295714 2.064597e-01
## factor(cyl)6 -3.03134449 1.40728351 -2.154040 4.068272e-02
## factor(cyl)8 -2.16367532 2.28425172 -0.947214 3.522509e-01
## hp           -0.03210943 0.01369257 -2.345025 2.693461e-02
## wt           -2.49682942 0.88558779 -2.819404 9.081408e-03
```

Since `fitstep` model gives a better explanation of variation and it has proper attributes we will choose it to be our final model. We see that for the given set of cars we find the correlation between transmission and MPG (with manual transmission resulting in better mileage), however we cannot prove its statistical significance. This happens due to the fact that the variables are highly correlated with each other and the choice of cars in the dataset does not represent a good coverage of all the variety of cars or a segment of the car market. Another model `fitp1` demonstrates significance of transmission for MPG but it omits important variables that are proved to have much influence on mileage, so we have to reject it in favour of `fitstep` to avoid biased results.

Appendix

Figure 1. Miles per Gallon by Transmission

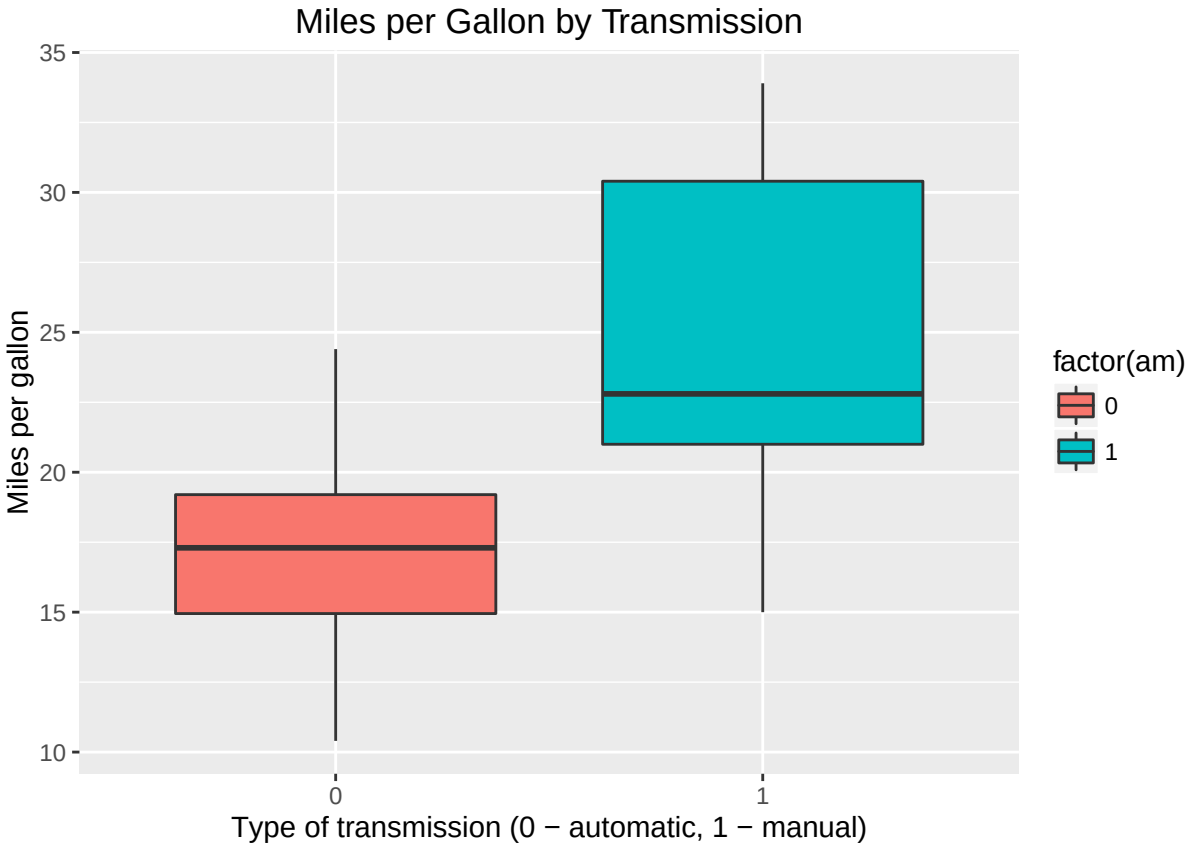

Figure 2. Correlation Matrix Plot

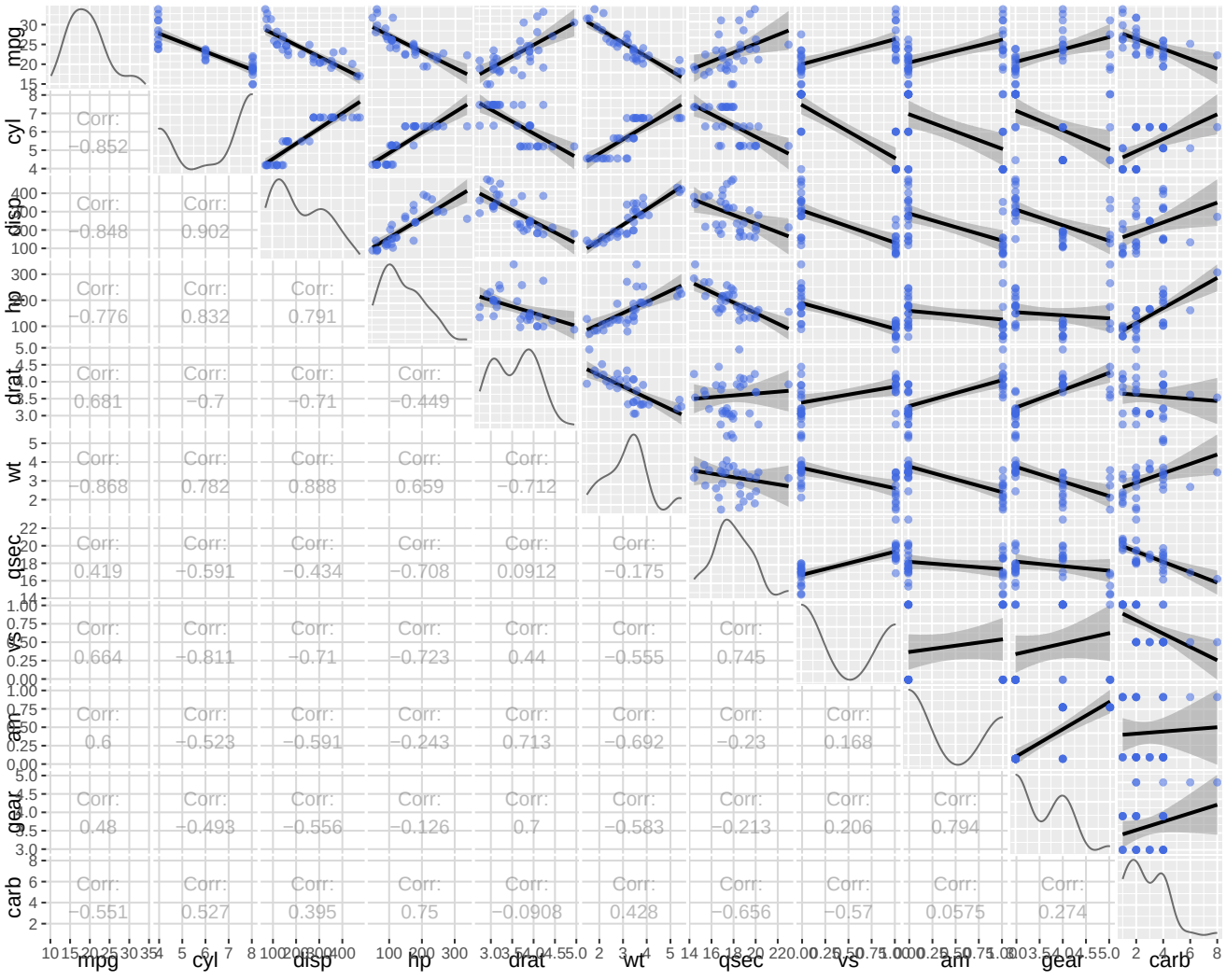

Figure 3. fitp1 Model Diagnostic Plots

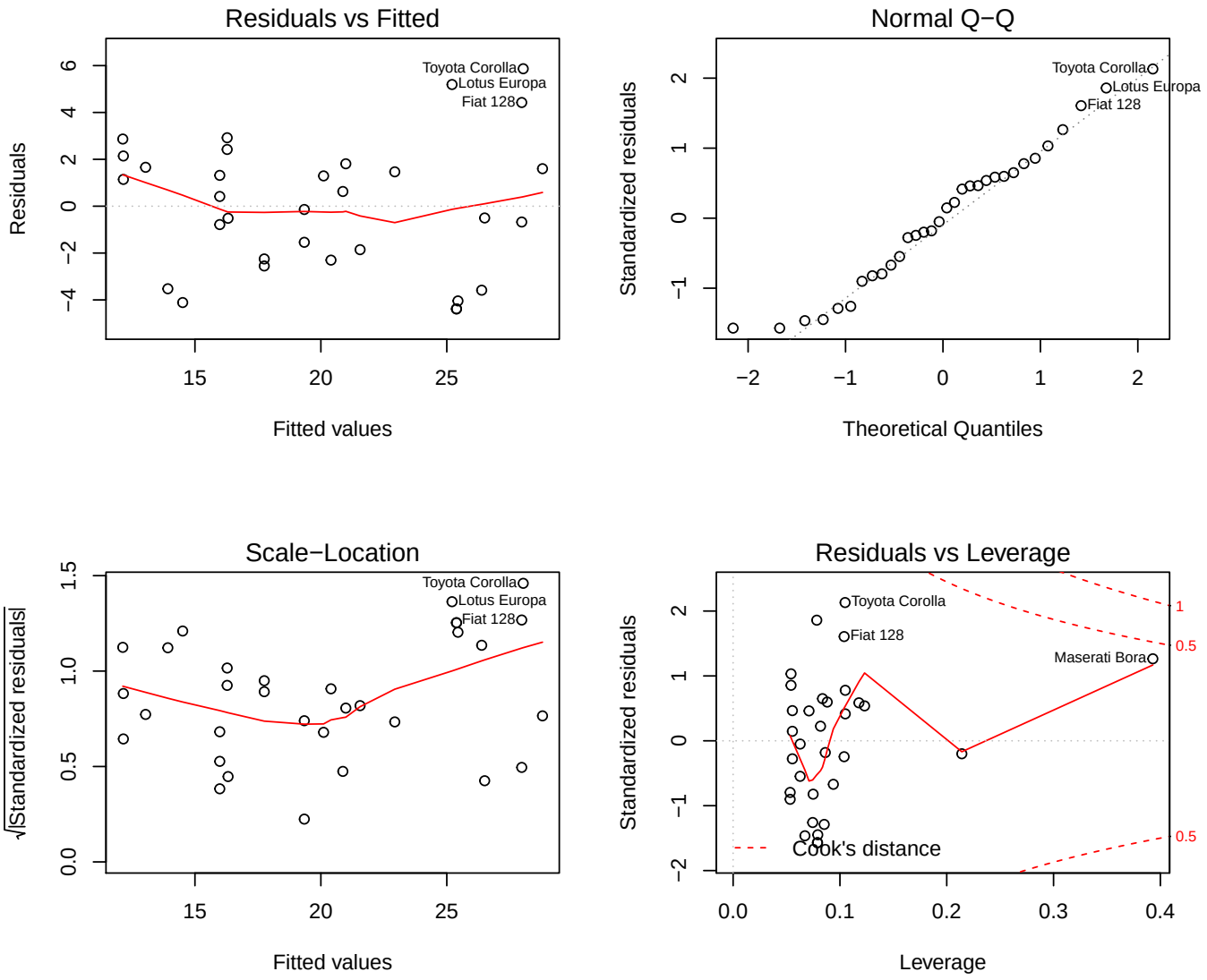

Figure 4. fitp2 Model Diagnostic Plots

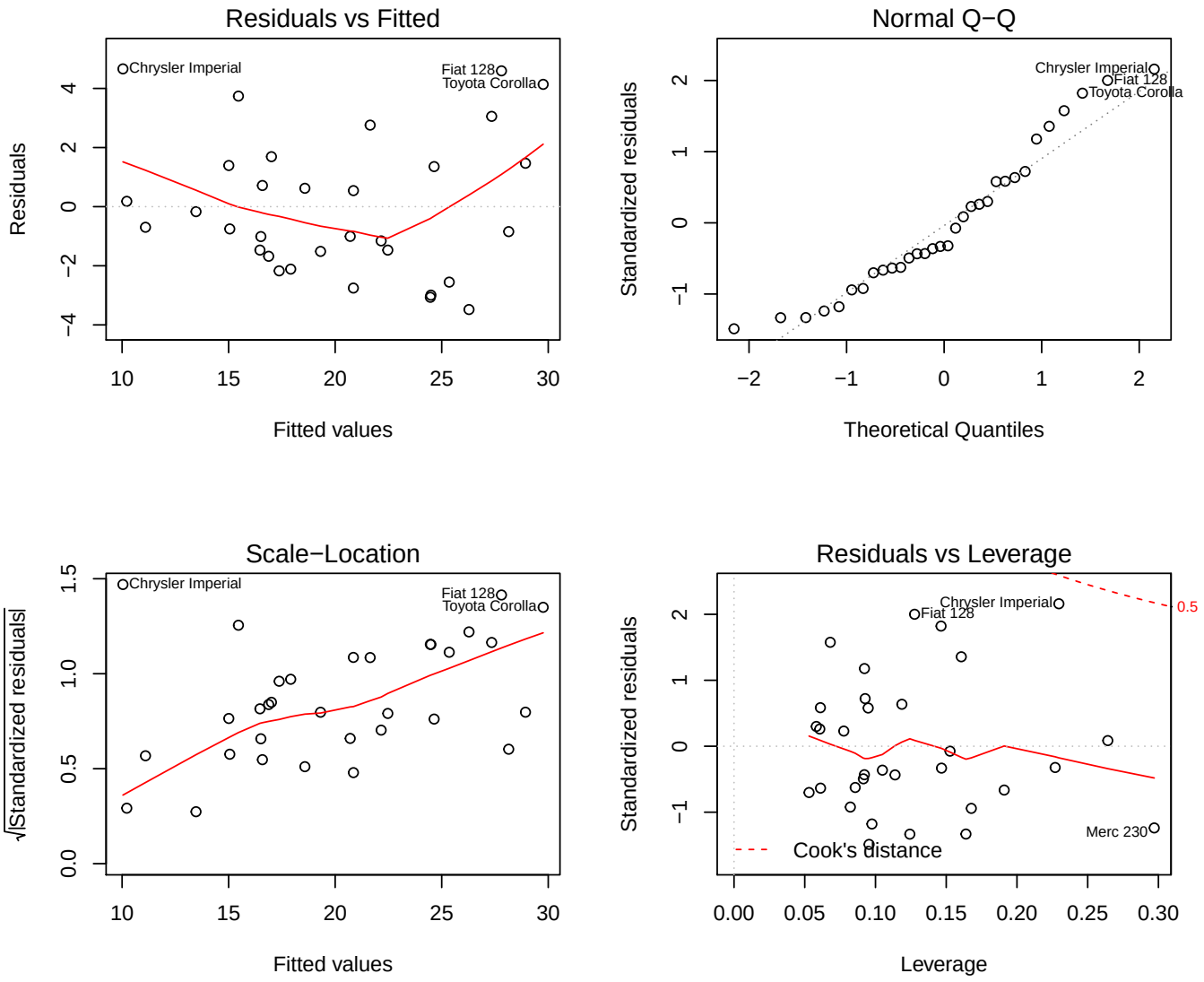

Figure 5. Final Multivariable Linear Regression Model (fitstep) Diagnostic Plots